A study of stock returns based on ARMA-GARCH model: Taking the CSI 300 Index as an example

Yating Chen[1,*]

[1]School of Economics & Management, Nanjing University of Science & Technology, Nanjing, China

*Corresponding author: 1274841030@qq.com

Abstract. This paper selects the daily return data of CSI 300 index from January 2, 2014 to September 30, 2020 to study the characteristics of CSI 300 index return volatility. An ARMA(3,3) model was fitted to the daily log returns of the CSI 300 index using R software, and the ARCH effect was found to exist. The returns were then fitted with an ARMA(3,3)-GARCH(1,1) model, the model fitting effect was tested, and finally the short-term returns of the CSI 300 index were predicted. The results of the empirical analysis show that the CSI 300 index return series has characteristics such as non-normal distributivity, spikes and thick tails, and aggregation.

Keywords: CSI 300 index; ARCH effect; ARMA-GARCH model.

1. Introduction

Time series analysis is an important branch of statistics discipline and is one of the important tools in the field of economic forecasting research, describing the pattern of historical data over time and used to forecast economic data. With the development of economic globalisation and financial internationalisation, the stock market as an important part of the global financial market has attracted much attention. The volatility characteristics of the stock market is an important subject of research by scholars at home and abroad, and more and more investors are aware of the importance of stock market forecasting.

In the real stock market, macroeconomic developments, the enactment of laws and regulations, investor expectations, company fundamentals, black swan events and so on can all affect stock market sentiment, and time series data on stock prices or returns often exhibit stochasticity. The use of time series analysis models for empirical analysis to forecast stock price or return trends provides investors and financial regulation with some basis for decision making. With the continuous development of mathematical theoretical research, data analysis tools and data mining techniques, scholars have adopted various different methods and tools to analyse financial time series and propose various stock price forecasting models.

The CSI 300 Index covers nearly 60% of the market capitalisation of Shanghai and Shenzhen, and has good market representativeness, becoming one of the guideposts for observing the whole stock market in China. This paper selects the CSI 300 Index as the research object and establishes a model for empirical analysis.

2. Review of the literature

In recent years, foreign scholars have conducted a large number of studies on stock price volatility, mainly focusing on the theory of stock price volatility and the methods and models for measuring volatility. The most common phenomenon in the stock market is volatility. Mandlebrot (1963), through empirical research, found that price fluctuations in the stock market show a certain degree of agglomeration, and his findings laid the theoretical foundation for later studies on the characteristics of stock price volatility.

In his analysis of the UK inflation series, Engle (1982) found a certain pattern in the volatility of the yield series, i.e. the agglomeration effect of volatility, on the basis of which an autoregressive conditional heteroskedastic (ARCH) model was developed. heteroskedastic) or ARCH model for short. [1] However, the ARCH model is only applicable to the short-term autocorrelation process of the heteroskedastic function. In practice, some residual series of the heteroskedastic function is with long-term autocorrelation, in order to correct this problem, Bollerslev proposed the generalized autoregressive conditional heteroskedastic model (GARCH model for short). [2]

The GARCH model can well solve the problem of lag order in the model, simplify the parameter estimation process, and use the significance of the GARCH term, ARCH term and the coefficient of the constant term in its conditional variance equation to determine the agglomeration of volatility. [3] However, the GARCH model cannot effectively explain the asymmetric effects of time series volatility. engle (1993) extended the ARCH model in order to describe the leverage effect of stock price volatility and carried out empirical analysis on the GARCH, EGARCH , TGARCH and other models.Zhao Guojian and Liu Jing (2010) studied stock market volatility and conducted empirical analysis using a GARCH model, and concluded that the TGARCH model could better fit the relative return series of the SSE index. [4] Jiang Xiangcheng and Xiong Yamin (2017) empirically analyzed the volatility of China's stock market and concluded that the EGARCH model could better fit the time series of daily return volatility of Shanghai and Shenzhen markets, and there was significant asymmetry in China's stock market. [5] Li Yannan (2018) established a GARCH model to empirically analyze the daily closing price of the Shanghai Composite Index and concluded that the logarithmic price of the Shanghai Stock Exchange Index has the phenomenon of volatility aggregation. [6]

A large number of studies at home and abroad have shown that GARCH models are more suitable for studying the volatility agglomeration of financial time series, while TGARCH and EGARCH models work well in terms of asymmetry of financial time series volatility. Liu Huizhen investigates the extent to which GARCH, EGARCH and TGARCH models characterise the volatility of broad market index returns in the Chinese stock market and finds that under the assumption of conditional t-distribution, both EGARCH and TGARCH models can better measure the asymmetry of the impact of good and bad news on return volatility, while GARCH models cannot explain the asymmetry of the impact of good and bad news of the same size shocks on volatility asymmetrically. [7]

With the continuous development of mathematical theoretical research, data analysis tools and data mining techniques, scholars have adopted various methods and tools to analyse financial time series and propose various stock price prediction models.

3. Method

3.1 ARMA model and GARCH model

ARIMA(p,d,q) model, where AR(p) is the autoregressive of order p, d denotes the order of the difference made to change the time series from non-stationary to stationary, and MA(q) is the moving average of order q. The essence of the ARIMA model is a combination of the difference operation and the ARMA model, where any non-stationary series can be achieved by differencing of the appropriate order to achieve post-differential stationarity, and an ARMA model can be fitted to the post-differential series ARMA model fitting is performed.

The autoregressive moving average ARMA(p,q) model has both a p-order autoregressive and a q-order moving average, and is structured as follows：

$$\phi_1 \mathrm{X}_{\mathrm{t-1}} + \phi_2 \mathrm{X}_{t-2} + \ldots + \phi_p \mathrm{X}_{t-p} + \varepsilon_t = \mathrm{X}_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \tag{1}$$

Specially then, known as the centralised ARMA(p,q) model.In equation (1), it can be seen that AR(p), MA(q) is a special form of ARMA(p,q) corresponding to q=0,p=0 respectively, these three models are only suitable for describing smooth time series, while the actual application is often non-smooth time series, for non-smooth time-series data sometimes carry out one or two differences can be transformed into a smooth time-series data.

If the assumption of variance homogeneity does not hold, heteroskedasticity may occur.As the sample data is smooth, this paper establishes an ARMA(p,q) model of the return series to describe the volatility characteristics of the CSI 300 index, conducts ARCH effect tests and solves the heteroskedasticity problem with a GARCH model.

The GARCH model adds to the ARCH model by considering the p-order autocorrelation of the heteroskedasticity function, which can effectively fit a heteroskedasticity function with long term memory, and the Garch model considers both the autoregressive and moving average components of the heteroskedasticity.

The GARCH model proposed by Bollerslev attributes present volatility to past moment volatility and past moment error, so the GARCH model can explain the clustering phenomenon of financial return volatility. However, the error term at past moments in the GARCH model is a squared term, which only measures the effect of the size of the message and does not reflect the different effects of good and bad news.

4. Empirical Study

4.1 Sample Selection and Data Sources

The CSI 300 Index sample covers nearly 60% of the market capitalisation of Shanghai and Shenzhen, which has good market representativeness and becomes one of the indicators for observing the whole stock market in China. This paper selects the CSI 300 Index as the research object to study the volatility of the Chinese stock market. The data in this paper comes from the Resset database, and the daily closing price data of the CSI 300 Index from 2nd January 2014 to 30th September 2020 is selected, with a total of 1647 sample data. In this paper, the logarithmic return is used in calculating the daily return.

4.2 Descriptive statistics of sample data

Descriptive statistics were conducted using R software to plot the time series of daily returns of the CSI 300 index, as shown in Figure 4.1.

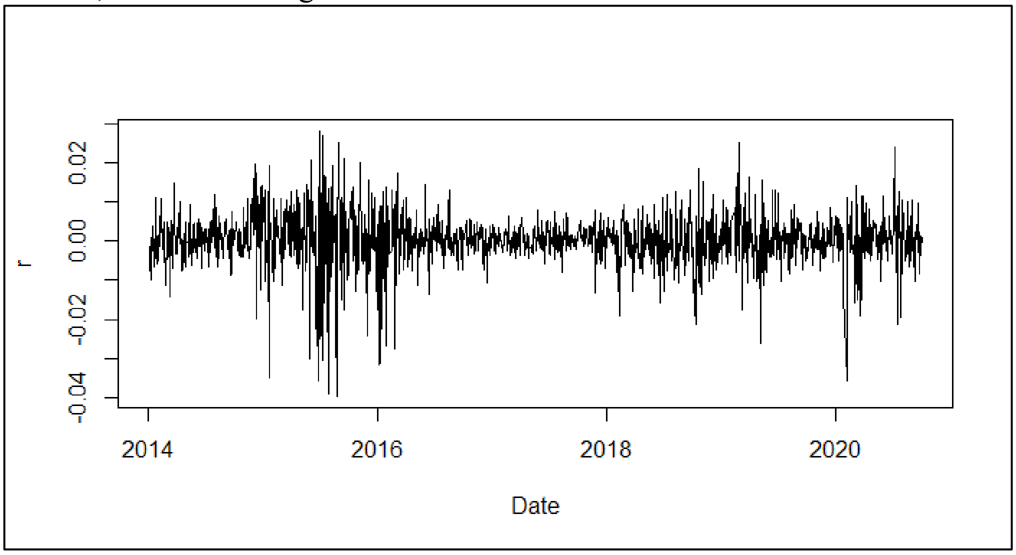

Fig.1 Time-series chart of CSI 300 Index returns

The time series chart portrays the overall volatility characteristics of the CSI 300 Index from 2nd January 2014 to 30th June 2020. As shown in Fig.1, the CSI 300 Index daily return series from June 2015 to March 2016, from August 2018 to May 2019 and from March to July 2020 are more volatile and the rest of the time periods are more volatile with small fluctuations in the daily return of the index in some time periods with persistently high volatility and persistently low volatility in some time periods.

Table 1. Descriptive statistical analysis of yields

Observations	Mean	Median	Std.Dev	Skewness	Kurtosis	Jarque-Bera	Prob.
1647	0.000179	0.000289	0.006558	-0.938130	6.206807	947.272	0.0000

From Table 1, it can be seen that the mean value of the CSI 300 daily return is 0.000179, the standard deviation is 0.006558, the skewness is -0.93813 less than 0, the kurtosis is 6.206807 much greater than 3, the distribution of the return is more left-skewed than the normal distribution, from the time series chart of the CSI 300 daily return can be seen that the return has the distribution characteristics of a sharp peak and thick tail. The JB statistic is used to test the distribution characteristics of the data. If the p-value is smaller, it means that the more deviation from normal distribution combined with the p-value of JB test in Fig.2 and Table 1, the original hypothesis that daily returns obey normal distribution is rejected.The Normal Q-Q diagram is plotted in the R language environment, as shown in Fig.2.

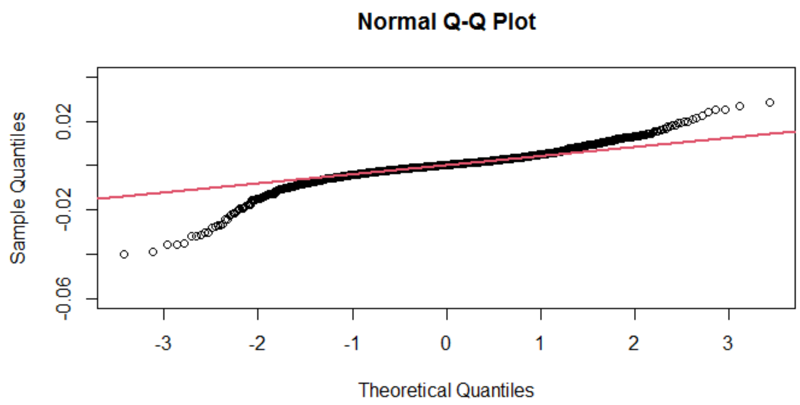

Fig.2 Normal Q-Q chart of CSI 300 daily returns

4.3 Robustness test

When performing statistical tests and model estimation on time series data, it is often required that the series be stationary, and both ARMA and GARCH models require stationary time series data. Non-stationary data can lead to invalid statistical results and inaccurate model estimates, so an ADF test is required to test the stationarity of the return series.

The ADF unit root test was performed using R software. From the test results, it can be seen that ADF=-11.604, P=0.01, so the original hypothesis of unit root exists in the return series at 5% significance level is rejected.The daily return series of CSI 300 index is smooth time-series data. According to Fig.1 time series of returns, the daily returns of the CSI 300 Index fluctuate up and down around the horizontal axis with no significant time trend.

4.4 Autocorrelation test

Financial time series data are often non-linear and non-stationary in nature, especially stock prices and stock returns which are often affected by their own and perturbation terms in the earlier and later periods and are autocorrelated. When autocorrelation exists in the sample data, the ordinary least squares estimation OLS method is an unbiased estimate but the model results are no longer valid. The serial correlation of the sample data should be examined before modelling the time series data for analysis. In the R software, the autocorrelation and partial autocorrelation plots of the return series are plotted, as shown in Fig.3.

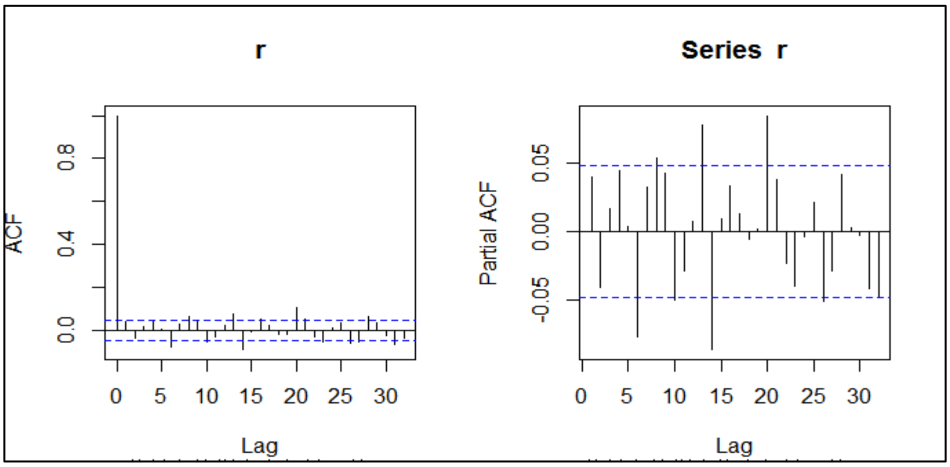

Fig.3 ACF and partial ACF plots

The two horizontal lines in the ACF plot represent the upper and lower bounds of the 2 standard errors.From the ACF and partial ACF plots, it can be seen that the autocorrelation coefficients and partial autocorrelation coefficients are trailing and are suitable for the ARMA(p,q) model, but the lag order p,q is more difficult to determine. In this paper, the Ljung-Box statistic of the daily return series of the CSI 300 index is used to determine whether the daily return data are autocorrelated. The LB statistic is obtained through the Box.test function in R software, Q(4)=18.269, and the corresponding p-value is significantly less than α=0.01, then the original hypothesis of no serial correlation is rejected,there is autocorrelation in the daily return data of the CSI 300 index.

4.5 ARMA model

An attempt was made to fit the series using an ARMA(p,q) model to eliminate serial autocorrelation. In this paper, great likelihood estimation is used to estimate the parameters of the mean equation. as there are many possible combinations of the order p, q of the model, this paper fixes the order of the ARMA(p,q) model according to the AIC information criterion. the function of the AIC criterion is defined as:

$$\mathrm{AIC}(p, q, \mu) = \log\sigma_{\varepsilon}(p, q, \mu) + 2(p + q + 1)/N \tag{2}$$

AIC is based on the theory of entropy, the complexity and fit of the model is judged, assuming there are n models to choose from, by calculating the AIC value of each model, the model that achieves the minimum AIC value is selected. in R software, by calling the auto.arima function to fix the order of the yield series, according to the AIC criterion, R software selects the ARMA(3,3) model to fit the yield sequence, obtaining the model calibre as:

$$r_t = 0.5157r_{t-1} - 1.0036r_{t-2} + 0.3713r_{t-3} + \varepsilon_t - 0.4755\varepsilon_{t-1} + 0.965\varepsilon_{t-2} - 0.2894\varepsilon_{t-3} \tag{3}$$

The parameter estimation results of the mean equation when p=3,q=3 are shown in Table 2. From Table 2, it can be seen that the coefficients in the ARMA(3,3) model are significant, and then the residual series of the fitted model are tested for white noise.

Table 2. Estimated parameters of the ARMA(3,3) model

Models	Estimated values	Error	T-test	$Pr(>\lvert t\rvert)$	Significance
AR(1)	0.5157	0.2672	1.9300	0.0426	**
AR(2)	-1.0036	0.0365	-27.4959	0.0000	***
AR(3)	0.3713	0.2242	1.6556	0.0980	*
MA(1)	-0.4755	0.2745	-1.7322	0.0781	*
MA(2)	0.9650	0.0451	21.4444	0.0000	***
MA(3)	-0.2894	0.1579	-1.8324	0.0671	*

Note: *** is significant at 1% confidence level, ** is significant at 5% confidence level, * is significant at 10% confidence level, - is insignificant.

From the test results, it can be seen that the p-values are significantly greater than 0.05 and the original hypothesis that the residual series is a white noise series cannot be rejected. The residual series is a white noise series and the model has extracted the level information in the return series. For the complete analysis of the observation series to focus on both level and volatility, the ARMA model extracts the information related to level, and the following analysis of the information related to volatility of returns.

4.6 ARCH effect tests and GARCH modelling

After the mean equation of the model was obtained using the maximum likelihood estimation, correlation tests and conditional heteroskedasticity tests were done on the squared residual series of returns. Residual squared time series plots, autocorrelation plots of residual squares and partial autocorrelation plots of the model fit using R are shown in Fig.4 and Fig.5.

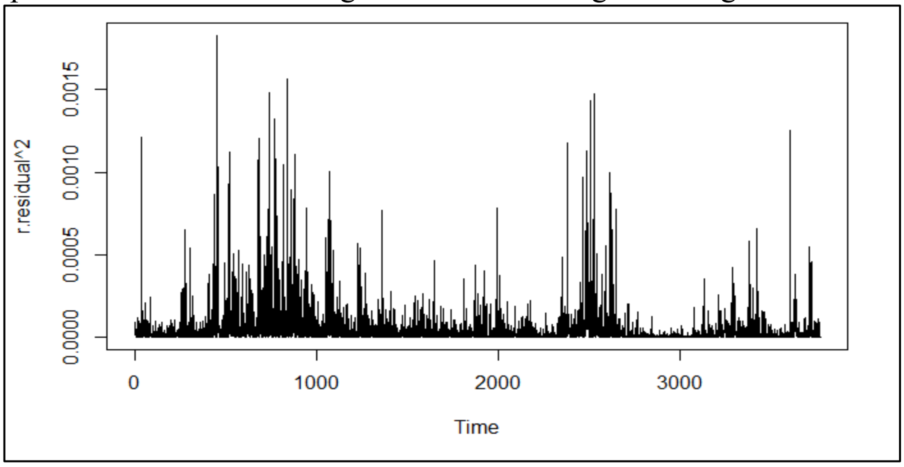

Fig.4 Residual squared time series plots

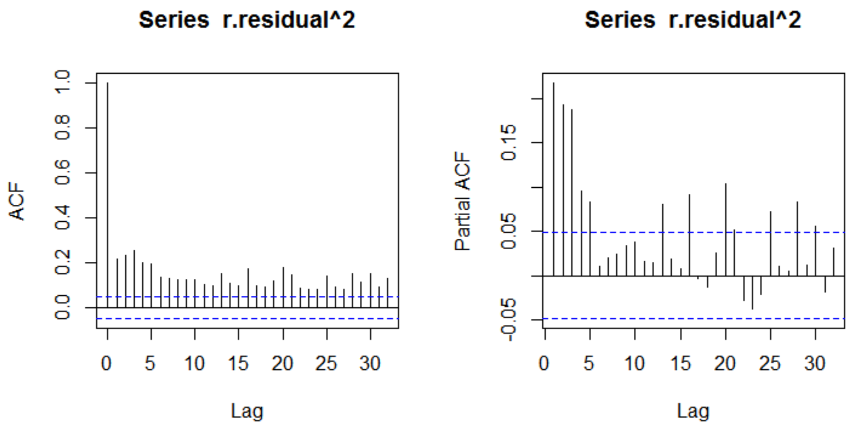

Fig.5 ACF and partial ACF plots of squared residuals

From Fig.4 it can be seen that the residual series shows significant heteroskedasticity and from Fig.5 it can be seen that the residual squared series is correlated.The ARCH-LM test is constructed with the idea that if the residual series is non-trivial and has a clustering effect, then the residual squared series is usually autocorrelated, and an ARCH(q) model can be used to try to fit the residual squared series. In this paper, the ARCH-LM test is carried out using R software, taking lags of order 1 to 5.

The results of the LM test show that the p-values for lags 1-5 are significantly less than the 5% confidence level, the series is significantly non-square and there is an ARCH effect, and the ARCH model can be used to extract the correlation information contained in the residual squared series. However, the ARCH model essentially uses the qth order moving average of the residual squared series to fit the current period heteroskedasticity function values and is only applicable to the short term autocorrelation process of the heteroskedasticity function.

GARCH model was chosen to extract volatility information in this paper. The GARCH (1,1) model was fitted in the R software.The estimated results of the parameters in the GARCH (1,1) model are shown in Table 3, and the coefficients are all significant as can be seen from Table 3.

Table 3. Results of GARCH(1,1) parameter estimation

Parameters	Estimated values	Error	T-test	$Pr(>\|t\|)$	Significance
a0	2.268e-07	6.496e-08	3.491	0.000481	***
a1	0.08459	0.005605	15.094	0.000000	***
b1	0.9151	0.004965	184.327	0.000000	***

Note: *** is significant at 1% confidence level.

The fitted GARCH(1,1) model is:

$$\begin{cases} \varepsilon_t = \sqrt{h_t} e_t \\ h_t = 0.9151 h_{t-1} + 0.08459 \varepsilon_{t-1}^2 \end{cases} \tag{4}$$

The residual series generated by the GARCH(1,1) model fit were then tested for correlation and found to be uncorrelated, with no ARCH effect in the residual series and a better model fit.

Combining the horizontal model and the volatility model, the complete fitted model was obtained as:

$$\begin{cases} r_t = 0.5157 r_{t-1} - 1.0036 r_{t-2} + 0.3713 r_{t-3} + \varepsilon_t - 0.4755 \varepsilon_{t-1} + 0.965 \varepsilon_{t-2} - 0.2894 \varepsilon_{t-3} \\ \varepsilon_t = \sqrt{h_t} e_t \\ h_t + 0.9151 h_{t-1} + 0.08459 \varepsilon_{t-1}^2 \end{cases} \tag{5}$$

5. Summary

In this paper, the daily closing price of CSI 300 index from January 2nd, 2014 to September 30th, 2020 was selected to study the daily return and volatility of CSI 300 index. Using R software, the ARMA(3,3) model was determined using the AIC criterion, and the ARCH effect was found by testing the residual term, so the GARCH model was introduced and the ARMA(3,3)-GARCH(1,1) model was finally established to fit the returns. The results of the empirical analysis show that the CSI 300 index return series exhibits the characteristics of spikes and thick tails and asymmetry. The volatility of the CSI 300 Index shows a clustering effect, when the volatility of the market is high, the high volatility tends to last for a period of time, and then the stock market enters a less volatile phase.

The model used to fit yields in this paper is a model commonly used in traditional financial time series analysis, which is not very effective in fitting yield series in the context of increasingly complex financial time series data. There are currently many scholars in China who combine wavelet analysis theory and GARCH models to empirically analyse the volatility characteristics and impact factors of China's stock market yields. The traditional GARCH model assumes the same response of conditional variance to positive and negative shocks, but in practice in the stock market, the impact of good news and bad news on the stock market is not the same GARCH model cannot explain the asymmetry of the impact of good and bad news shocks of the same size on volatility.

References

[1] Engle R F . Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation[J]. Econometrica, 1981, 50(4):987-1008.

[2] Baillie R T , Bollerslev T , Mikkelsen H O. Fractionally integrated generalized autoregressive conditional heteroskedasticity[J]. Journal of Econometrics, 1996, 74.

[3] Wang ZhengLong. Time series analysis [M]. China Statistics Press, 2000.

[4] Zhao Guojian, Liu Jing. A study of Shanghai stock market volatility based on GARCH model[J]. Enterprise Herald, 2010(12):2.

[5] Jiang Xiangcheng;. A study on the volatility of China's stock market based on GARCH family model[J]. Journal of Southwest Normal University: Natural Science Edition, 2017, 42(2):5.

[6] Li Yannan. A study of China's stock price volatility based on GARCH model - taking SSE index as an example[J]. China Business Journal, 2018(15):2.

[7] Liu, Huizhen, Zhang, Chenghu, Li, Yulin. Forecasting financial asset return volatility - a study based on the jumping behavior of China's stock market[J]. Contemporary Economic Science, 2011(5):10.